HW\_Thinking\_Week3\_Saturaday

Xiang Guo (郭想)

**Thinking 1 如何使用用户标签来指导业务（如何提升业务）**

简单来说，标签能够在一下三种场景发挥作用来提升业务。

1. 获客。标签能够帮助产品或者服务拉到更多的客户，其原因就在于标签能够相对精准的定位到那些可能被这款产品吸引的消费群体，给他们推荐感兴趣的内容和服务，这样新客户的获取概率就大大提升。
2. 黏客。对于刚刚上手使用产品的客户来说，标签能够更多更准确地推荐给用户一些个性化的内容，这样一来客户就更容易被这些个性化推荐所吸引并且持续性地使用该产品。
3. 留客。任何一款产品和服务都会有顾客流失，但是在使用标签之后就能更好的对这些用户流失做一个分析，比如根据标签判断哪一类群体更容易离开，然后分析原因为什么这一类人会对产品失去兴趣。然后就可以在未来做出改进来留住这一部分用户。

**Thinking 2 如果给你一堆用户数据，没有打标签。你该如何处理（如何打标签）**

对于没有打标签的用户数据，最常用的方式就是使用聚类算法来对特征降维把相似的用户数据归到一类，然后再根据这一类的特点人工的打上标签。

**Thinking 3 准确率和精确率有何不同（评估指标）**

准确率和精准率的计算方式不同。

准确率的公式为TP+TN）/(P+N)，这个很容易理解，就是被分对的样本数除以所有的样本数，通常来说，正确率越高，分类器越好；

精确率的公式为TP/（TP+FP），也就是说在所有被分为正例的示例中实际为正例的比例。和准确率一样，也是越高约好，说明分类器很少会把不是正例的判断为正例。

但是我们知道对于某个具体的分类器而言，我们不可能同时提高这些指标，很多时候我们想要某一个指标提高就不得不牺牲一些其他的指标。所以到底希望哪一个高还是取决于这个分类器使用的场景，也就是使用者更在意哪一个指标。比如说一个判断作弊的系统，它可能准确率很高，但是主要是因为对于没有作弊的情况判断准确，但是对于判断为有作弊的情况，有很多被归到了没有作弊的例子（这个例子中Positive为没有作弊）。这样的系统就可能漏过很多作弊的例子。

那么这种情况下可以调整模型比如稍微牺牲一些准确率或者其他指标但是提升精确率，来确保虽然有些没有作弊的被判断为了作弊，但是绝大部分作弊的都保证被准确地抓到。

**Thinking 4 如果你使用大众点评，想要给某个餐厅打标签。这时系统可以自动提示一些标签，你会如何设计（标签推荐）**

一般有三种方法：

1. 推荐大众点评上最热门的标签。
2. 推荐这一类餐厅最热门的标签。
3. 推荐用户自己经常使用的标签。

**Thinking 5 我们今天使用了10种方式来解MNIST，这些方法有何不同？你还有其他方法来解决MNIST识别问题么（分类方法）**

|  |  |  |
| --- | --- | --- |
|  | 优点 | 缺点 |
| KNN | 对低维数据效果比较好 | 消耗内存，因为需要储存所有的实列。 |
| 逻辑回归 | 模型相对来说比较简单，抗噪声干扰能力比较强，而且因为它的输出是几率，可解释性比较好。 | 需要线性或者近似线性的特征。 |
| SVM | 不需要线性特征，而且对于高维数据的分类效果比较好比如文本 | 训练比较低效。而且可解释性比较低。没有标准的方法处理多类的问题，基本是一个二值分类器。 |
| 朴素贝叶斯 | 分类效率稳定，对小规模的数据表现很好。对缺失值也不太敏感。 | 属性个数比较多或者属性之间相关性较大时，分类效果不好。需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测效果不佳 |
| 决策树 | 决策树主要的优点在于不需要线性的特征或者相互之间有线性作用的特征。而且可以很好的处理高纬度空间的大数量训练集。模型本身也比较容易解释。 | 不好做模型维护。而且有可能会过拟合或者陷入局部最小值的情况。 |
| 神经网络 | 神经网络的好处在于获得的数据越多，模型效果越好。所以在数据越来越多的当下，表现也就越来越好，一般被用于自然语言处理或者计算机视觉 | 训练需要很长的时间，也需要很多数据。对算力的要求比较高。而且是个黑盒子，也就是可解释性比较差。 |

**Thinking 5 使用TPOT等AutoML工具，又怎样的好处和不足（ppt里面是这一道题）**

使用AutoML的好处是在进行数据清洗之后，AutoML就可以自动来进行特征选择，预处理和模型选择，然后使用训练数据来对模型进行训练。好处就在于省去了我们特征选择和创建，然后训练模型这些步骤。而且TPOT可以自动测试多个模型的结果然后把一个表现较好的模型返回出来。

不足在于处理大数据的时候会比较慢，对算力的要求比较高。然后目前只支持监督学习，也就是数据集必须带有标签。